岡山理科大学紀要 第43号 A pp. 57-65(2007)

# マルチプルアラインメントに対する Maximum Weight Trace 法のクリークを考慮した貪欲法

西野　史芳・片山　謙吾*・南原　英生*・成久　洋之*
岡山理科大学大学院工学研究科情報工学専攻
*岡山理科大学工学部情報工学科

(2007年10月1日受付、2007年11月2日受理)

## 1. まえがき

本研究はマルチプルアラインメントを対象とし，Maximum Weight Trace 法を用い，貪欲法を改良した手法を提案する．

アラインメントとは，バイオインフォマティクスにおいて，複数の DNA の塩基配列やタンパク質のアミノ酸配列を並置し，相同性を求める手法である[1]．相同性とは，複数の生物の共通祖先に由来する子孫間の類似性である．つまり，アラインメントとは共通祖先から分岐し，次第に進化していった複数の生物の共通点を見つけ出す手法であるといえる．特に，本研究で対象とするマルチプルアラインメントは 3 本以上の配列によるアラインメントである．バイオインフォマティクスが発達する以前，タンパク質の構造や機能の決定はタンパク質に対する直接的な実験により行われてきた．しかし，タンパク質に対する直接的な実験によって，タンパク質の構造や機能を決定するよりも，このタンパク質に対応する DNA の配列を情報処理技術を用いて解析する方がはるかに易しい．そこで，情報処理技術を用いたアラインメントが重要な手法となっている[2]．

マルチプルアラインメントに対する代表的な手法としては，厳密解法である動的計画法を複数の配列に対し，部分的に繰り返し適用させる ClustalW[4] や，メタヒューリスティックアルゴリズムである遺伝的アルゴリズムを用いる SAGA[5] などが知られている．これら ClustalW や SAGA など既存の手法は，配列に対し，局所的な操作によって解を求めている．しかし局所的な操作では，配列が大規模になればなるほど，全域的に最適な解を求めにくくなる．そこで，より広範囲に配列の関係を考慮できる Maximum Weight Trace 法[6] が提案された．Maximum Weight Trace 法は，アラインメントをアラインメントグラフとして表現する手法である．このアラインメントグラフにより，全ての配列の情報を一度に表現することが可能である．アラインメントグラフに基づくマルチプルアラインメントは，貪欲法によって簡潔にアラインメントを得ることができる．しかし貪欲法では，全域的な情報に関係なく解を求めることになる．そこで本研究ではアラインメントグラフにおけるクリークに着目する．アラインメントグラフに対し貪欲法を適用させる際に，クリークを考慮することで単純な貪欲法に比べ，より全域的に良好な解が得られると考えられる．このクリークを考慮した貪欲法の性能を検証するため，他手法と比較実験を行った．

## 2. アラインメント

アミノ酸は 20 種類の文字からなる配列で表される．そして，複数の配列の相同性を求めるために整列させる操作がアラインメントである．特にペアワイズアラインメントについては Needleman-Wunsch 法によって最適なアラインメントを得ることができる．

### 2.1 アミノ酸配列

タンパク質はアミノ酸からなる化合物である．このアミノ酸は 20 種類あり，各アミノ酸は，以下に示すように 1 文字のアルファベットの 20 種類で表される．

A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y

したがって，あるタンパク質は1列のアミノ酸配列として表現される．このアミノ酸配列は生物が進化していく過程で，あるアミノ酸が他のアミノ酸と入れ替わる置換，新たにアミノ酸が加わる挿入，すでにあるアミノ酸が失われる欠失が行われる．しかし，進化していく過程の中で変化しないアミノ酸もある．そして，複数のタンパク質が同じ構造と機能を持っていれば相同性を保持しているという．

### 2.2 アラインメント

アラインメント (図1) とは，複数の配列を縦に比較し，文字が一致するように，隙間を挿入し整列させる方法である．ここで，$\Sigma$ を隙間を表す ' - ' 以外の各アミノ酸を表す有限個の記号体系，' - ' を含む有限個の記号体系を $\hat{\Sigma} = \Sigma \cup \{\text{`}-\text{'}\}$ とする．また各配列を $S_1, \ldots, S_n$，各配列の長さを $l_1, \ldots, l_n$ とし，$S_n$ に含まれる $l_n$ 個の各文字を $s_{n,1}, s_{n,2}, \ldots, s_{n,l_n}$ とする．このとき，$S_1, \ldots, S_n$ のアラインメント $A$ は $n$ 個の文字列 $\hat{S}_1, \ldots, \hat{S}_n \in \hat{\Sigma}$ からなる $n \times l$ 次元配列 $A = (a_{i,j})$ となる．この $A$ は以下の特徴を持つ．

- $a_{i,j} \in \hat{\Sigma} \quad \forall 1 \geq i \geq n, 1 \geq j \geq n$
- $\hat{S}_i = S_i \setminus \{\text{`}-\text{'}\}$
- 縦列に隙間がない場合は $\max\{l_1, ..., l_n\} \leq n \leq \Sigma_{i=1}^{k} n_i$

$A$ を得るための操作は $0 \leq a < n$，$0 \leq b < n$，$a \neq b$，$0 \leq \alpha < l_a$，$0 \leq \beta < l_b$ のとき

- $s_{a,l_\alpha}, s_{b,l_\beta} \in \Sigma$ かつ $s_{a,l_\alpha} \neq s_{b,l_\beta}$ の場合の置換
- $s_{a,l_\alpha} =$ ' $-$ ' かつ $s_{b,l_\beta} \in \Sigma$ の場合の挿入
- $s_{a,l_\alpha} \in \Sigma$ かつ $s_{b,l_\beta} =$ ' $-$ ' の場合の欠失

である．このアラインメントは $n = 2$ の場合をペアワイズアラインメント，$n \geq 3$ の場合をマルチプルアラインメントと呼ぶ．

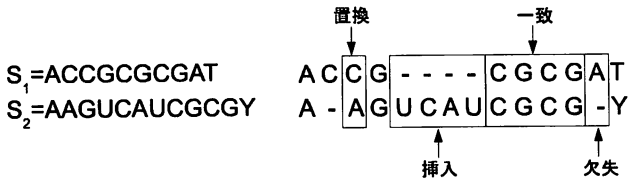


図1 アラインメントの例

### 2.3 アラインメントスコア

最適なアラインメント $A$ を得るためには，$A$ を評価しなければならない．そのための得点をアライメントスコアと呼ぶ．このアラインメントスコアを得るときは，スコア行列 (図2)[7] を用いる．スコア行列は，置換されやすいものほど得点が高く，置換されにくいものほど得点が低く表されている．このスコア行列により得られた得点に対し，隙間の長さに応じてペナルティを与えることで，アラインメントスコアを表す．スコア行列を $m$，アラインメントスコアを $sc$，隙間の数を $g$，ペナルティを $p$ とすると，あるペアワイズアラインメント $A$ におけるアラインメントスコア $sc$ は

$$sc(A) = -(g \times p) + \Sigma_n m(\hat{S}_{1,l_n}, \hat{S}_{2,l_n})$$

となる．

### 2.4 Needlman-Wunsch 法

ペアワイズアラインメントは動的計画法を用いることで，最適なアラインメントを得ることができる．この動的計画法を用いることで最適なアラインメント $A$ を得る方法を Needlman-Wunsch 法[8] と呼ぶ．2本

| | G | A | S | T | P | L | I | M | V | D | N | E | Q | F | Y | W | K | R | H | C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G | 5 | | | | | | | | | | | | | | | | | | | |
| A | 1 | 2 | | | | | | | | | | | | | | | | | | |
| S | 1 | 1 | 2 | | | | | | | | | | | | | | | | | |
| T | 0 | 1 | 1 | 3 | | | | | | | | | | | | | | | | |
| P | -1 | 1 | 1 | 0 | 6 | | | | | | | | | | | | | | | |
| L | -4 | -2 | -3 | -2 | -3 | 6 | | | | | | | | | | | | | | |
| I | -3 | -1 | -1 | 0 | -2 | 2 | 5 | | | | | | | | | | | | | |
| M | -3 | -1 | -2 | -1 | -2 | 4 | 2 | 6 | | | | | | | | | | | | |
| V | -1 | 0 | -1 | 0 | -1 | 2 | 4 | 2 | 4 | | | | | | | | | | | |
| D | 1 | 0 | 0 | 0 | -1 | -4 | -2 | -3 | -2 | 4 | | | | | | | | | | |
| N | 0 | 0 | 1 | 0 | -1 | -3 | -2 | -2 | -2 | 2 | 2 | | | | | | | | | |
| E | 0 | 0 | 0 | 0 | -1 | -3 | -2 | -2 | -2 | 3 | 1 | 4 | | | | | | | | |
| Q | -1 | 0 | -1 | -1 | 0 | -2 | -2 | -1 | -2 | 2 | 1 | 2 | 4 | | | | | | | |
| F | -5 | -4 | -3 | -3 | -5 | 2 | 1 | 0 | -1 | -6 | -4 | -5 | -5 | 9 | | | | | | |
| Y | -5 | -3 | -3 | -3 | -5 | -1 | -1 | -2 | -2 | -4 | -2 | -4 | -4 | 7 | 10 | | | | | |
| W | -7 | -6 | -2 | -5 | -6 | -2 | -5 | -4 | -6 | -7 | -4 | -7 | -5 | 0 | 0 | 17 | | | | |
| K | -2 | -1 | 0 | 0 | -1 | -3 | -2 | 0 | -2 | 0 | 1 | 0 | 1 | -5 | -4 | -3 | 5 | | | |
| R | -3 | -2 | 0 | -1 | 0 | -3 | -2 | 0 | -2 | -1 | 0 | -1 | 1 | -4 | -4 | 2 | 3 | 6 | | |
| H | -2 | -1 | -1 | -1 | 0 | -2 | -2 | -2 | -2 | 1 | 2 | 1 | 3 | -2 | 0 | -3 | 0 | 2 | 6 | |
| C | -3 | -2 | 0 | -2 | -3 | -6 | -2 | -5 | -2 | -5 | -4 | -5 | -5 | -4 | 0 | -8 | -5 | -4 | -3 | 12 |

図 2　スコア行列の例

の配列 $S_a$，$S_b$ について，最適な $A$ を得るため，これらの配列をもとに $l_a \times l_b$ の 2 次元配列 $D_{i,j}$ を

$$
\begin{aligned}
D_{0,0} &= 0 \\
D_{0,j} &= \Sigma_{n=1}^{j} m(-, s_{b,l_n}) \\
D_{i,0} &= \Sigma_{n=1}^{i} m(s_{a,l_n}, -) \\
D_{i,j} &= \max \left\{ \begin{array}{l} D_{i,j-1} + m(-, s_{b,l_j}) \\ D_{i-1,j-1} + m(s_{a,l_i}, s_{b,l_j}) \\ D_{i-1,j} + m(s_{a,l_i}, -) \end{array} \right\} \quad \forall i, j > 0
\end{aligned}
$$

と定義する．このとき $D_{i,j}$ =max となるのは，部分配列 $s_{a,l_n}, s_{b,l_n}$ の一致または置換，挿入，欠失からなる各組合せのうち，最も得点の高い組合せ，すなわち最適なアラインメントスコアとなる組合せである．したがって，最適な $A$ を得るには，最も高いアラインメントスコアとなる部分配列を通る経路を選択すればよい．

最適な $A$ は，トレース行列を用いることで容易に手に入れられる．配列 $S_a$，$S_b$ に対するトレース行列 $tr_{i,j}$ は，要素 $tr_{i,j} \subseteq \{\leftarrow, \uparrow, \nwarrow\}$ からなる行列と定義される．各要素は

$$
\begin{aligned}
tr_{0,0} &= \emptyset \\
tr_{0,j} &= \leftarrow \\
tr_{i,0} &= \uparrow \\
\nwarrow \in tr_{i,j} &\iff D_{i,j} = D_{i-1,j-1} + m(s_{a,l_i}, s_{b,l_j}) \quad \forall i, j > 0 \\
\uparrow \in tr_{i,j} &\iff D_{i,j} = D_{i-1,j} + m(s_{a,l_i}, -) \\
\leftarrow \in tr_{i,j} &\iff D_{i,j} = D_{i,j-1} + m(-, s_{b,l_j})
\end{aligned}
$$

となる．この $tr_{i,j}$ の右下角 $tr_{l_a,l_b}$ から左上角 $tr_{0,0}$ にトレースバックする (図 3)．例えば，$i$ 番目の $\nwarrow$ もしくは $\uparrow$ を $s_{a,l_i}$ に一致または置換とし，すべての $\leftarrow$ を挿入とすることで，$\hat{S}_a$ を得る．同様に，$j$ 番目の $\leftarrow$ もしくは $\nwarrow$ を $s_{b,l_j}$ に一致または置換とし，すべての $\uparrow$ を欠失とすることで，$\hat{S}_b$ を得る．

## 3. Muximum Weight Trace 法

Maximum Weight Trace 法はマルチプルアラインメントをグラフとして表す．しかし，このグラフから最適なアラインメントを得ることは NP 困難である[6]．そのため，最適なアラインメントを得るために必要

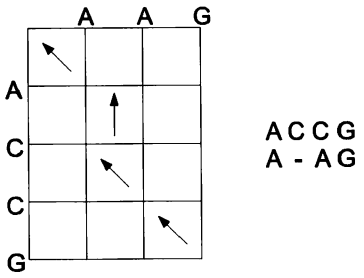


図3　トレースバックにより得られるアラインメントの例

な情報のみからなるグラフを作る.

### 3.1 アラインメントグラフ

マルチプルアラインメントにおいて，$n$ 個の配列の各文字を頂点とみなすと，頂点集合 $V$ および各頂点間を結ぶ枝 $e$ の集合 $E$ からなる完全 $n$ 部グラフ $G=(E,V)$ として表すことができる．そして，各枝には非負の重み $w(e)$ がある．このとき $V$ および $E$ は

$$\begin{aligned} V &= s_{i,p} \quad \forall i=1,\ldots,n \quad \forall p=1,\ldots,l_i \\ E &= \{(s_{i,p},s_{j,q})|1\le i<j\le n, 1\le p\le l_i, 1\le q\le l_j\} \end{aligned}$$

である．このように表されるグラフをアラインメントグラフ (図4) と呼ぶ.

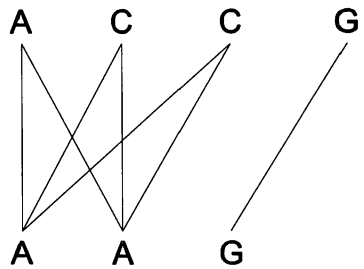


図4　アラインメントグラフの例

### 3.2 Maximum Weight Trace 問題

ここであるアラインメント $A$ を考える．この $A$ によって表現される枝の全集合を $A$ のトレース $T\subset E$ とする．そして，$T$ の重みは $\Sigma_{e\in T}w(e)$ となる．このとき完全アラインメントグラフを元に $T\subset E$ から，最大の重みとなる $T$ を選ぶことを Complete Maximum Weight Trace 問題と呼ぶ．同様に，アラインメントグラフを元に $T\subset E$ から，最大の重みとなる $T$ を選ぶことを Maximum Weight Trace 問題と呼ぶ.

### 3.3 アラインメントグラフの作成

Maximum Weight Trace 問題は NP 困難である．そこで，枝集合 $E$ をアラインメントとなる可能性の高いものに限定することを考える．その解決策として，Needlman-Wunsch 法によるペアワイズアラインメントを $n$ 個の配列の全ての組合せについて行うことで枝を得る.

また，各枝の重みは Sequence Identity Score($SIS$) によって求める．$SIS$ はペアワイズアラインメントを行い，その中で文字が一致した割合を表す.

$$SIS=\frac{m_1}{m_1+m_0}\cdot 100$$

ここで，$m_1$ は文字が一致した数，$m_0$ は文字が不一致であった数を表す.

### 3.4 ペアワイズアラインメントの拡張

ペアワイズアラインメントでは局所的に最適なアラインメントを得ることしかできない．そこで，各アラインメント全体の全域的な情報をアラインメントグラフに反映させるため，ペアワイズアラインメントの拡張を行う．ペアワイズアラインメントの拡張は，一致または置換された文字のペアと，そのペアとは異なる一致または置換された文字のペアについての推移律を元に行う[11]．

このペアワイズアラインメントの拡張により得られたアラインメントグラフ上の各配列 $S_{i_1},\ldots,S_{i_k}$ について，$k \geq 2$ であるとき，各配列に含まれる文字を $s_{i_\alpha,p_\alpha}, 1 \geq \alpha \geq k$ とする．このとき，$\forall\alpha = 1,\ldots,k-1$ であり，かつ $s_{i_\alpha,p_\alpha}$ と $s_{i_{\alpha+1},p_{\alpha+1}}$ の間に枝が存在するとき，文字 $s_{i_1,p_1}$ から文字 $s_{i_k,p_k}$ への枝を edge path と呼ぶ．そして，文字 $s_{i_1,p_1}$ から文字 $s_{i_k,p_k}$ の間に枝があるとき，文字 $s_{i_1,p_1}$ と文字 $s_{i_k,p_k}$ の間の枝をレベル $k$ の transitive edge(図 5) と呼ぶ．

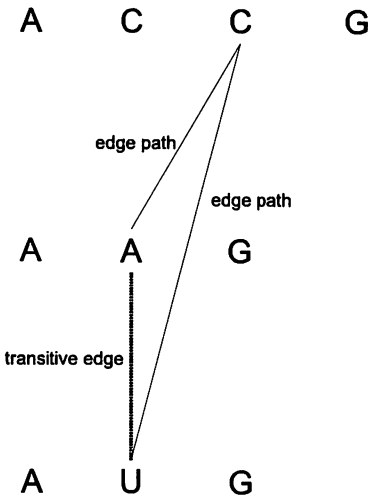


図 5　transitive edge の例

### 3.5 拡張アラインメントグラフ

これまでに得られたアラインメントグラフを元にマルチプルアラインメントを得る際，トレースをグラフ理論的に特徴付ける必要がある[6]．そこで，拡張アラインメントグラフ (図 6) を用いる．拡張アラインメントグラフでは新たに枝 $H = \{(s_{i,p}, s_{i,p+1}) | 1 \leq i \leq n, 1 \leq p \leq l_i - 1\}$ を定義し，矢印で表す．この枝 $H$ の重みは 0 である．このとき拡張アラインメントグラフは $\overline{G} = (V, E, H)$ と表される．

このグラフ $\overline{G}$ では配列の頂点 $v_1,\ldots,v_n, n \geq 2$ について，$1 \leq i < n$ であり，かつ $(v_i, v_{i+1}) \in E$ または $(v_i, v_{i+1}) \in H$ とする．ここで，もし最初の頂点と最後の頂点が同じであるならば，このパスを cycle と呼ぶ．また，$\overline{G} = (V, E, H)$ が拡張アラインメントグラフであるならば，トレース $T$ によって導かれた $T \subseteq E$ かつ $\overline{G}^T = (V, T, H)$ は拡張アラインメントグラフである．このとき，矢印を含まない $\overline{G}^T$ が全て cycle であったならば，$T$ はトレース可能であると呼ぶ．

## 4. クリークを考慮した貪欲法

Maximum Weight Trace 問題は貪欲法によってアラインメントを得ることができる．しかし，貪欲法では最適なアラインメントが得られるとは限らない．そこで本研究では，貪欲法を用いる際にクリークの重みを考慮する．クリークの重みを考慮することで，よりよいアラインメントとなる枝が選ばれやすくなると考えられる．

### 4.1 貪欲法

Muximum Weight Trace 問題は貪欲法を用いることで，簡潔にアラインメントを得ることができる．この貪欲法ではアラインメントグラフに含まれる枝集合 $E$ を重みについてソートし，重みの大きいトレース

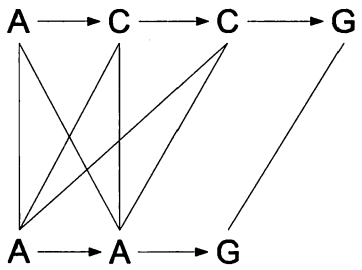


図 6　拡張アラインメントグラフの例

可能な枝 $e$ から順にトレースに追加していく．そして，アラインメントグラフからトレース可能な枝が無くなったときに終了する．

しかし，この方法では全域的に最適なアラインメントになるとは限らない．例えば，ある $e_a$ をトレースに追加して構成される部分配列よりも，$e_a$ を追加することで，トレースに追加可能ではなくなった枝の部分集合 $E*\backslash e_a$ を追加した場合の部分配列の方がよりよいアラインメントとなる場合がある．以下において，重みが最大となるクリークを考慮する貪欲法を示す．

### 4.2　クリークを考慮した貪欲法

本法では，アラインメントグラフ $\overline{G}$ に含まれる枝集合 $E$ を重み $w(e)$ についてソートし，$w(e)$ の大きいトレース可能な $e$ から順にトレース $T$ に追加しようとするところまでは同じである．この $e$ を $T$ に追加する操作のとき，$e$ を $T$ に追加することで，$T$ にクリークが作られる場合は，そのクリークの重みよりも重くなるクリークを形成する $e$ を探す．そして，トレース可能な枝のなかから，クリークの重みを最大にする $e$ を $T$ に追加する．このときの手順を以下に示す．

1. トレース可能な $e$ のうち，最も重み $w(e)$ の大きい $e$ をトレース $T$ とする．
2. 1 のとき，$T$ 中の枝を含むクリークに，非トレースの頂点を含む $e$ を加えるとき，この頂点と同じ配列の中で構成されるクリークの重みが最大になる $e$ を選択する．
3. アラインメントグラフにトレース可能な $e$ がなければ終了．さもなければ 1 へ戻る．

## 5. 実験結果

クリークを考慮した貪欲法を評価するために貪欲法及び ClustalW，SAGA と比較した．問題例はマルチプルアラインメントのベンチマークである BAliBASE[12] の Reference1 より 76 問を使用した．また，評価値として用いてるスコアは，BAliBASE におけるアラインメントスコア算出プログラム baliscore で求めた値である．この baliscore では最適なアラインメントのスコアを 1 とし，悪いアラインメントほど 0 に近づく．表 1 に各手法の実験結果を示す．

表 1 より，クリークを考慮した貪欲法は単純な貪欲法に比べ，改善される場合と改悪される場合は半々であることが分かる．ここで改善または改悪された場合の値に着目すると，最も改善された 1tgxA ではスコアが 0.93 増加しているのに対し，最も改悪された 2cba ではスコアが 0.30 しか減少していない．このことから，改善と改悪の例題数は同じでも，スコア上では改善される幅が大きいといえる．

また，貪欲法，クリークを考慮した貪欲法共に他の比較手法に比べ，優れている例題は極めて少ない．しかし，クリークを考慮した貪欲法は貪欲法の一種であり，貪欲法は解を構築するための解法である．したがって，クリークを考慮した貪欲法を局所探索法などの解改善法と組み合わせることで，より良いアラインメントが得られると考えられる．

## 6. むすび

本研究で示した手法は，貪欲法を基本とし，ある枝を加える際に，クリークの重みが最大になるものを選択した．しかし，提案法はアラインメントグラフ全体に比べ，局所的にクリークを選択する．ゆえに今後の

表 1　貪欲法，クリークを考慮した貪欲法，ClustalW，SAGA の実験結果

| | 貪欲法 | | Clique 貪欲法 | | CluatalW | | SAGA | |
|---|---|---|---|---|---|---|---|---|
| 問題例 | スコア | 時間 | スコア | 時間 | スコア | 時間 | スコア | 時間 |
| 1aab | 0.619 | 0.02 | 0.616 | 0.11 | 0.909 | 0.01 | 0.839 | 4 |
| 1aboA | 0.289 | 0.03 | 0.304 | 0.13 | 0.690 | 0.01 | 0.521 | 19 |
| 1aho | 0.824 | 0.03 | 0.796 | 0.08 | 0.894 | 0.01 | 0.960 | 8 |
| 1csp | 0.952 | 0.03 | 0.956 | 0.18 | 0.987 | 0.01 | 0.955 | 5 |
| 1csy | 0.746 | 0.06 | 0.745 | 0.19 | 0.931 | 0.02 | 0.888 | 5 |
| 1dox | 0.924 | 0.04 | 0.924 | 0.17 | 0.917 | 0.01 | 0.864 | 10 |
| 1fjlA | 0.828 | 0.04 | 0.829 | 0.18 | 0.997 | 0.01 | 0.991 | 7 |
| 1fkj | 0.878 | 0.06 | 0.878 | 0.43 | 0.941 | 0.03 | 0.954 | 10 |
| 1fmb | 0.861 | 0.03 | 0.863 | 0.18 | 0.973 | 0.01 | 0.972 | 5 |
| 1hfh | 0.769 | 0.07 | 0.758 | 0.55 | 0.869 | 0.04 | 0.903 | 21 |
| 1hpi | 0.600 | 0.02 | 0.603 | 0.11 | 0.863 | 0.01 | 0.901 | 6 |
| 1idy | 0.375 | 0.01 | 0.379 | 0.11 | 0.626 | 0.01 | 0.348 | 6 |
| 1krn | 0.840 | 0.04 | 0.843 | 0.10 | 0.988 | 0.01 | 0.981 | 9 |
| 1pfc | 0.731 | 0.06 | 0.728 | 0.23 | 0.881 | 0.03 | 0.913 | 17 |
| 1plc | 0.871 | 0.05 | 0.871 | 0.16 | 0.924 | 0.02 | 0.951 | 10 |
| 1r69 | 0.196 | 0.02 | 0.196 | 0.09 | 0.510 | 0.01 | 0.563 | 7 |
| 1tgxA | 0.651 | 0.01 | 0.744 | 0.07 | 0.820 | 0.01 | 0.760 | 5 |
| 1tvxA | 0.163 | 0.01 | 0.163 | 0.03 | 0.331 | 0.01 | 0.456 | 8 |
| 1ycc | 0.716 | 0.04 | 0.725 | 0.19 | 0.873 | 0.02 | 0.779 | 10 |
| 2mhr | 0.882 | 0.06 | 0.886 | 0.54 | 0.982 | 0.03 | 0.961 | 11 |
| 2trx | 0.479 | 0.03 | 0.461 | 0.15 | 0.707 | 0.01 | 0.685 | 8 |
| 3cyr | 0.689 | 0.04 | 0.690 | 0.19 | 0.764 | 0.02 | 0.849 | 14 |
| 451c | 0.517 | 0.04 | 0.505 | 0.12 | 0.637 | 0.02 | 0.978 | 12 |
| 9rnt | 0.845 | 0.05 | 0.870 | 0.40 | 0.656 | 0.02 | 0.978 | 20 |
| 1ad2 | 0.755 | 0.13 | 0.754 | 0.99 | 0.903 | 0.01 | 0.881 | 22 |
| 1amk | 0.918 | 0.29 | 0.917 | 1.66 | 0.984 | 0.13 | 0.978 | 32 |
| 1ar5A | 0.852 | 0.12 | 0.848 | 0.90 | 0.974 | 0.06 | 0.977 | 17 |
| 1aym3 | 0.785 | 0.16 | 0.811 | 1.42 | 0.948 | 0.08 | 0.915 | 41 |
| 1ezm | 0.929 | 0.48 | 0.923 | 6.97 | 0.965 | 0.21 | 0.926 | 119 |
| 1havA | 0.131 | 0.17 | 0.138 | 1.27 | 0.334 | 0.12 | 0.311 | 64 |
| 1ldg | 0.814 | 0.29 | 0.844 | 3.12 | 0.942 | 0.15 | 0.938 | 31 |
| 1led | 0.461 | 0.16 | 0.442 | 1.49 | 0.923 | 0.09 | 0.834 | 28 |
| 1mrj | 0.868 | 0.19 | 0.853 | 1.92 | 0.927 | 0.11 | 0.923 | 63 |
| 1pgtA | 0.798 | 0.13 | 0.753 | 1.01 | 0.971 | 0.07 | 0.924 | 22 |
| 1pii | 0.666 | 0.18 | 0.664 | 1.81 | 0.826 | 0.11 | 0.848 | 83 |
| 1ppn | 0.599 | 0.22 | 0.597 | 1.11 | 0.988 | 0.11 | 0.985 | 50 |
| 1pysA | 0.917 | 0.19 | 0.925 | 1.89 | 0.936 | 0.11 | 0.922 | 38 |
| 1sbp | 0.358 | 0.32 | 0.370 | 3.72 | 0.571 | 0.18 | 0.440 | 85 |
| 1thm | 0.901 | 0.21 | 0.893 | 2.29 | 0.930 | 0.12 | 0.893 | 48 |
| 1tis | 0.917 | 0.37 | 0.920 | 2.32 | 0.971 | 0.20 | 0.947 | 103 |
| 1ton | 0.630 | 0.27 | 0.625 | 3.27 | 0.827 | 0.15 | 0.849 | 115 |
| 1uky | 0.182 | 0.12 | 0.185 | 0.93 | 0.628 | 0.09 | 0.443 | 62 |
| 1zin | 0.852 | 0.14 | 0.866 | 1.15 | 0.966 | 0.07 | 0.955 | 24 |
| 2cba | 0.640 | 0.28 | 0.610 | 1.48 | 0.814 | 0.17 | 0.831 | 63 |
| 2hsdA | 0.417 | 0.18 | 0.417 | 1.57 | 0.587 | 0.10 | 0.582 | 41 |
| 2pia | 0.504 | 0.19 | 0.476 | 1.71 | 0.757 | 0.13 | 0.623 | 94 |
| 3grs | 0.304 | 0.14 | 0.307 | 1.24 | 0.506 | 0.09 | 0.451 | 47 |
| 5ptp | 0.821 | 0.26 | 0.815 | 1.36 | 0.957 | 0.14 | 0.918 | 66 |
| kinase | 0.484 | 0.36 | 0.486 | 4.82 | 0.696 | 0.19 | 0.644 | 68 |
| 1ac5 | 0.696 | 0.56 | 0.689 | 7.69 | 0.878 | 0.36 | 0.798 | 326 |
| 1ad3 | 0.898 | 0.54 | 0.898 | 7.59 | 0.761 | 0.28 | 0.739 | 84 |
| 1adj | 0.858 | 0.48 | 0.865 | 6.67 | 0.947 | 0.24 | 0.956 | 42 |
| 1ajsA | 0.269 | 0.40 | 0.276 | 4.81 | 0.607 | 0.25 | 0.503 | 79 |
| 1cpt | 0.599 | 0.45 | 0.599 | 5.90 | 0.829 | 0.28 | 0.774 | 121 |
| 1dlc | 0.772 | 0.98 | 0.772 | 15.88 | 0.880 | 0.57 | 0.826 | 185 |
| 1eft | 0.812 | 0.40 | 0.811 | 5.12 | 0.887 | 0.24 | 0.891 | 83 |
| 1fieA | 0.846 | 1.35 | 0.845 | 23.67 | 0.918 | 0.66 | 0.846 | 138 |
| 1gowA | 0.580 | 0.64 | 0.568 | 9.01 | 0.896 | 0.42 | 0.714 | 82 |
| 1gpb | 0.930 | 3.12 | 0.930 | 32.83 | 0.984 | 1.43 | 0.962 | 398 |
| 1gtr | 0.855 | 0.89 | 0.854 | 16.78 | 0.965 | 0.41 | 0.936 | 118 |
| 1lcf | 0.880 | 3.32 | 0.881 | 54.99 | 0.790 | 1.51 | 0.723 | 705 |
| 1lvl | 0.880 | 3.32 | 0.881 | 55.58 | 0.475 | 0.35 | 0.395 | 122 |
| 1pamA | 0.118 | 1.21 | 0.113 | 19.64 | 0.610 | 0.88 | 0.414 | 318 |
| 1ped | 0.350 | 0.18 | 0.348 | 1.16 | 0.800 | 0.12 | 0.812 | 35 |
| 1pkm | 0.827 | 0.55 | 0.824 | 7.97 | 0.941 | 0.30 | 0.895 | 71 |
| 1rthA | 0.868 | 1.35 | 0.869 | 30.30 | 0.935 | 0.63 | 0.908 | 351 |
| 1sesA | 0.813 | 0.90 | 0.817 | 16.59 | 0.953 | 0.44 | 0.907 | 254 |
| 2myr | 0.074 | 0.50 | 0.079 | 5.06 | 0.388 | 0.44 | 0.185 | 200 |
| 3pmg | 0.924 | 0.87 | 0.923 | 14.16 | 0.973 | 0.47 | 0.937 | 224 |
| 4enl | 0.263 | 0.20 | 0.263 | 1.33 | 0.685 | 0.15 | 0.399 | 30 |
| actin | 0.905 | 0.70 | 0.904 | 12.80 | 0.948 | 0.34 | 0.917 | 160 |
| arp | 0.645 | 0.76 | 0.645 | 5.29 | 0.882 | 0.40 | 0.843 | 203 |
| gal4 | 0.242 | 0.65 | 0.239 | 10.11 | 0.463 | 0.37 | 0.403 | 168 |
| glg | 0.692 | 1.06 | 0.673 | 20.46 | 0.846 | 0.53 | 0.856 | 285 |

課題として，ある枝をトレースに加えようとするとき，さらにその先の枝を含めたクリークの重みが最大になる枝の集合を選択することが挙げられる．このことにより，より全域的に枝の集合の重みが大きくなると考えられる．

**参考文献**


1) D. W. Mount, "バイオインフォマティクス–ゲノム配列から機能解析へ–", メディカル・サイエンス・インターナショナル, 2002.

2) R. Durbin, S. R. Eddy, A. Krogh and G. Mitchison, "バイオインフォマティクス–確率モデルによる遺伝子配列解析–", 医学出版, 2001.

3) P. Clote and R. Backofen, "統計物理化学から学ぶバイオインフォマティクス", 共立出版, 2004

4) J. D. Thompson, D. G. Higgins and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice", Nucleic Acid Research, Vol. 22, 4673-4680, 1994.

5) C. Notredame and D.G. Higgins, "SAGA: Sequence Alignment by Genetic Algorithm", Nucleic Acid Research, Vol. 24, 1515-1524, 1996.

6) J.D. Kececioglu, "The Maximum Weight Trace Problem in Multiple Sequence Alignment", In Proceedings of the 4th Symposium on Combinatorial Pattern Matching, LNCS(684), 106-119, Springer, 1993.

7) M. O. Dayhoff, "Atlas of protein sequence and structure", National Biomedical Reserch foundation, Georgetown University, Washington, D.C., Vol.5, pp.89–99, 1972.

8) S. B. Needleman and C. D. Wunsch, "A General Method Applicable to the Search for Similatries in the Amino Acid Sequence of Two Proteins," J. Mol. Biol., Vol.48, pp.443–453, 1970.

9) G. Koller and G. R. Raidl, "An Evolutionary Algorithm for the Maximum Weight Trace Formulation of the Multiple Sequence Alignment Problem", Parallel Problem Solving from Nature, 302-311, 2004.

10) S. Leopold, "An Alignment Graph based Evolutionary Algorithm for the Multiple Sequence Alignment Problem", Master's thesis, Vienna University of Technology, Institute of Computer Graphics and Algorithms, February 2004.

11) C.Notredame, D. Higgins, and J. Heringa. "T-COFFEE: A nobel method for fast and accurate multiple sequence alignmnet", journal of Molecular Biology, 392, pp.205–217, 2000.

12) J. D. Thompson, F. Plewniak and O. Poch, "BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs", Bioinformatics, Vol.15, pp.87–88, 1999.

# A Greedy Method with Clique Finding for the Maximum Weight Trace Formulation of the Multiple Sequence Alignment Problem

Fumiyoshi NISHINO, Kengo KATAYAMA*,
Hideo MINAMIHARA* and Hiroyuki NARIHISA*

*Graduate School of Engineering,*
**Department of Information and Computer Engineering, Faculty of Engineering,*
*Okayama University of Science.*
*1-1 Ridai-cho, Okayama, 700-0005, Japan.*

(Received October 1, 2007; accepted November 2, 2007)

Most of the current techniques such as ClustalW and SAGA to the multiple alignment problem rely on local information to search near-optimal solutions. On the other hand, the maximum *weight trace* formulation can express global information. However the maximum weight trace problem is NP-hard. A simple technique to the multiple alignment problem with the maximum weight trace formulation is known to be greedy method. In this paper, we show a reformed greedy method that takes into account finding cliques. We compare the reformed method with the simple greedy method, ClustalW and SAGA on the benchmark set of multiple sequence alignment problem. The outcomes show that some results are improved by the greedy method with clique finding.

**Keywords:** multiple alignment; muximum weight trace formulation; greedy method; clique.